Volume 208, number 2 FEBS 4147 November 1986

cDNA and protein sequence of the NC1 domain of the α_2-chain of collagen IV and its comparison with α_1(IV)

Ulla Schwarz-Magdolen, Ilse Oberbäumer and Klaus Kühn*

Max-Planck Institut für Biochemie, D-8033 Martinsried, FRG

Received 11 September 1986

We present the complete cDNA and derived amino acid sequence of the non-collagenous domain NC1 of α_2(IV). Comparison with the corresponding NC1 domain of α_1(IV) reveals a high degree of homology at the protein level, in contrast to the barely homologous triple-helical sequences of both chains.

Basement membrane *Collagen IV* *NC1 domain* *Sequence homology*

1. INTRODUCTION

The main collagenous component of basement membranes is collagen IV. A molecule of collagen IV consists of two α_1- and one α_2-chain with about 1650 amino acid residues. Self-assembly of the molecules in the extracellular matrix occurs via like ends, e.g. four molecules overlap at the N-terminal triple helix (7 S domain) whereas two type IV monomers become connected via the C-terminal globular domain NC1. The network thus formed is stabilized by intermolecular disulfide bonds and non-reducible covalent crosslinks [1].

The primary structure of collagen IV is under investigation at the protein and cDNA level. For the α_1(IV)-chain, the N-terminal 7 S domain of human [2], the C-terminal NC1 domain of mouse and human and an adjacent portion of the triple helix [3–5] as well as 914 residues of the main triple-helical region of human [6] have been published. Some parts of this region are also known from murine and bovine [7,8] collagen IV. Of the α_2(IV)-chain, only the 511-residue-long C-terminal portion of the triple helix has been reported [9]. Here we present the primary structure of the adjacent non-collagenous globular domain NC1 of the α_2(IV)-chain as derived from cDNA sequence.

* To whom reprint requests should be addressed

2. MATERIALS AND METHODS

2.1. *Construction of the cDNA library and screening for α_2(IV)-specific clones*

The cDNA library cII was constructed with sucrose-gradient-purified RNA from PYS-2 cells [10] and the vector pUC931, using a simplified version of the procedure of Okayama and Berg [11]. pUC931 is a derivative of pUC9 [12] in which the polylinker of M13 tg131 [13] has been inserted. Details of the construction of this vector and the library cII will be published elsewhere [24]. Part of this library was screened with the nick-translated 3′ *Eco*RI/*Pst*I fragment of the α_2(IV) clone pAIIa. The clone pAIIa and the conditions for hybridization have been described by Schwarz et al. [9]. Screening of library cII identified clone pAIIc.

2.2. *Subcloning and sequencing*

Plasmid DNA of the clones was obtained by the method of Birnboim and Doly [14]. Relevant restriction fragments of clones pAIIa and pAIIc were subcloned into M13 tg130 and/or tg131 (Amersham) and sequenced by the chain termination method of Sanger et al. [15]. Each fragment was sequenced at least once on both strands.

2.3. *Computer analysis of sequences*

The computer program ALIGN [16] was used with the mutation data matrix, a penalty for a

Published by Elsevier Science Publishers B.V. (Biomedical Division)
00145793/86/$3.50

break of 12 and 100 random runs. The alignment score (in SD units) is the number of standard deviations by which the maximum score for the real sequences exceeds the average maximum score for random permutations of the sequences.

The program BESTFIT [17] produces an optimal alignment of sequences by scoring +1.0 for identical symbols, −0.9 for mismatches and additional negative values for introduced gaps. For the value ratio, the score is divided by the length (base pairs) of the shorter sequence. Thus the ratio is smaller than 1.0 for any two sequences that are not identical.

3. RESULTS

3.1. *Characterization of the clones*

We have isolated two overlapping cDNA clones, pAIIa and pAIIc, which together contain the sequence for the entire NC1 domain of the α_2(IV)-chain of mouse. Clone pAIIa has been isolated previously [9] and comprises about 560 bp coding for the NC1 domain, in addition to 1290 bp coding for the triple helix. Clone pAIIc starts at position 99 of the NC1 sequence, overlaps about 460 bp with the 3′-end of clone pAIIa and also includes the missing 3′ 130 bp of the NC1 domain and an additional 950 bp of the 3′-untranslated region. The restriction map of the two clones and the corresponding subclones in M13 are shown in fig.1, together with the sequencing strategy.

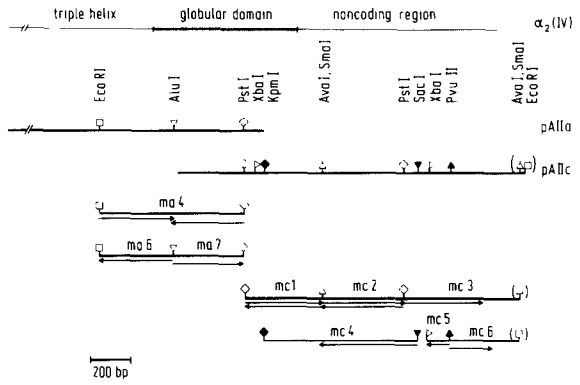

Fig.1. Restriction map and sequencing strategy of the α_2(IV)-specific cDNA clones pAIIa and pAIIc. The position of the two clones with respect to the α_2(IV) domains is given at the top. Segments ma and mc refer to subclones of pAIIa and pAIIc in the vectors M13 tg130 and/or tg131. Arrows indicate the direction of multiple sequencing. The restriction sites in brackets are contributed by the vectors.

3.2. *Characterization of the α_2(IV) NC1 domain at the protein level*

In fig.2, we present the complete cDNA and derived protein sequence of the α_2(IV) NC1 domain of mouse. The corresponding sequence for the α_1(IV)-chain of mouse and human has been published recently [3–5]. This allows for the first time a comparison of both chains in this important crosslinking domain. The α_2(IV) NC1 domain consists of 228 amino acid residues and is thus one residue shorter than the α_1(IV) sequence. Alignment of both protein sequences revealed an excellent homology (fig.2). Both domains contain 12 cysteine residues which occur at identical positions in a highly conserved surrounding, if two gaps in α_2(IV) (positions 92, 94) and one gap in α_1(IV) (position 176) are introduced. An alignment of the DNA sequence with the program BESTFIT (UWGCG, [17]) suggested some additional gaps after residue 197 which have been included in fig.2. 148 residues of both chains (65%) are identical; in 38 additional positions (16%), conservative amino acid substitutions have taken place, denoted by italic letters. Values for the homology at the protein and DNA levels are given in table 1.

Because of the overall homology of the NC1 domains of both chains, it could already be anticipated that the α_2(IV) NC1 domain exhibits a similar internal homology to that found for the α_1(IV)-chain [3–5]. The first and second halves can be aligned in such a way that all cysteine residues coincide and the neighbouring amino acid residues are identical or similar (program ALIGN [16]). The values for the internal homology of both chains are also given in table 1. This value is somewhat lower for the α_2-chain as more gaps are necessary for the alignment (not shown).

3.3. *Comparison at the DNA level*

Homology of two proteins at the protein level does not necessarily imply a high degree of homology at the DNA level. Comparison of the DNA sequences of α_1(IV) and α_2(IV) NC1 reveals many short stretches of 5–10 identical bases (cf. fig.2) interrupted by stretches where at least every

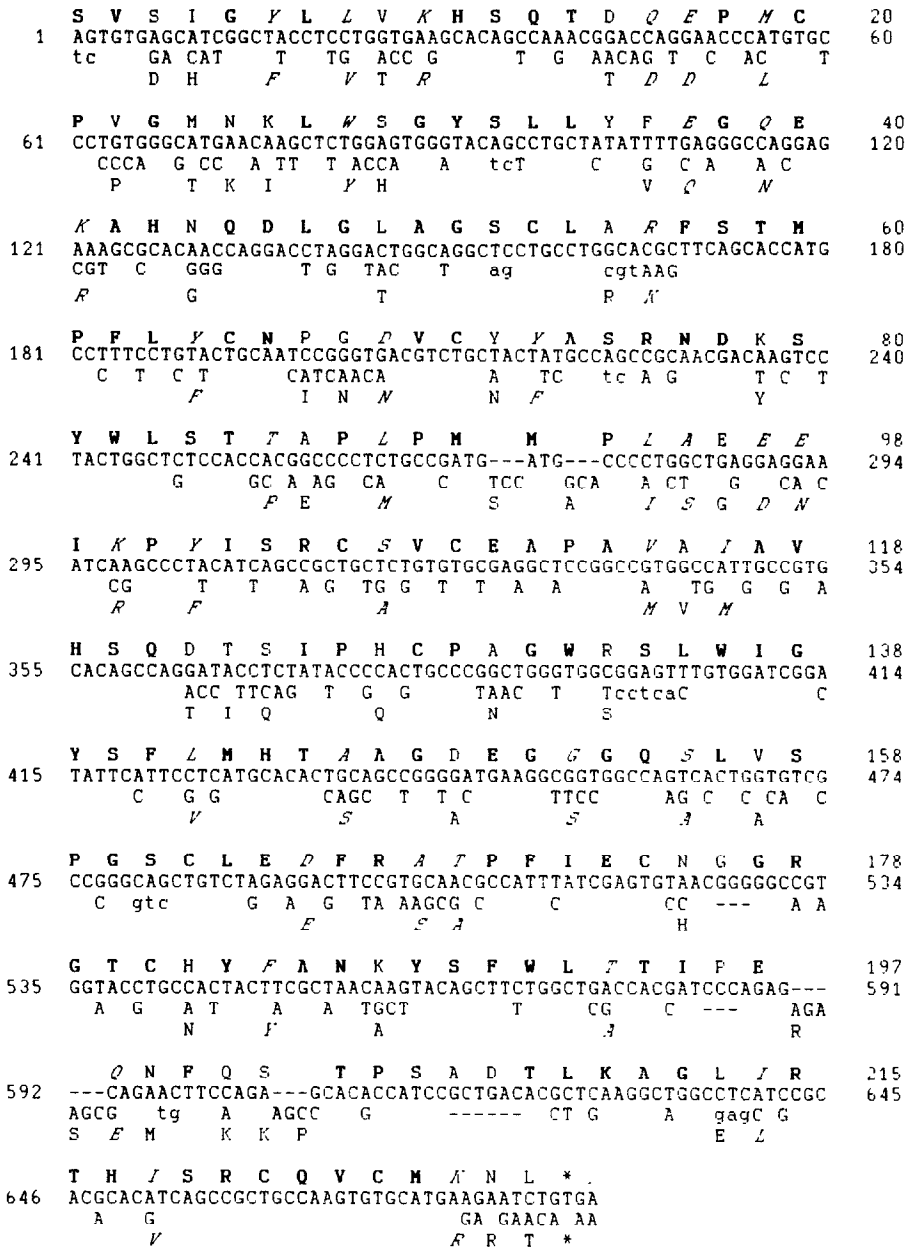

Fig.2. Comparison of cDNA and amino acid sequence of α_2(IV) NC1 of mouse with the corresponding sequences from the α_1(IV)-chain. The first row shows the protein sequence of α_2(IV) NC1 with its cDNA sequence in the second line. The third line gives those bases of the matched α_1 cDNA sequence which differ. The fourth row contains the derived amino acids that are not conserved. Bold letters indicate identical amino acids, italic letters conservative substitutions. Bases printed in lower-case letters are complementary to the sequences above. The first amino acid residue (Ser) of the α_2(IV) NC1 domain had been included in the triple helix previously [9].

Table 1

Homology between α_1 and α_2(IV)

Chains	Domain	Value
Protein level		
α_2(IV)-α_1(IV) (mouse)	NC1[a]	64.9%
α_2(IV) (mouse)-α_1(IV) (human)	NC1[a]	65.4%
α_1(IV) (mouse)-α_1(IV) (human)	NC1[a]	96.9%
α_1(I)-α_2(I) (chick)[b]	CPP[a]	61.2%
α_1(IV) (mouse) internal repeat	NC1[c]	18.1 SD
α_2(IV) (mouse) internal repeat	NC1[c]	16.1 SD
DNA level		
α_2(IV)-α_1(IV) (mouse)	NC1[d]	0.287
α_2(IV) (mouse)-α_1(IV) (human)[e]	NC1[d]	0.293
α_1(IV) (mouse)-α_1(IV) (human)[e]	NC1[d]	0.761
α_1(IV)-α_2(IV) (mouse)[f]	TR[d]	0.09
α_1(IV) (mouse)-α_1(IV) (human)[e,f]	TR[d]	0.647

[a] % of identical residues per number of residues of shorter chain
[b] Sequences taken from [18]
[c] Program ALIGN
[d] Program BESTFIT (ratio), gap weight 5, gap weight length 0.3
[e] Sequences taken from [4,5]
[f] Sequences taken from [3,9]

CPP, carboxypropeptide; TR, triple helix

third base is different. There are only two longer strings of matches (14 and 24 bases) at positions 169–182 and 660–684. These data suggest that under moderately stringent conditions no crosshybridization between both chains will occur with DNA probes from this area. This conclusion holds all the more for the triple-helical domain (cf. table 1 [9]).

Of the 148 identical amino acids, 72 (48.6%) are coded for by identical codons while 65 (44.2%) have one base exchanged. The remaining 11 codons with two or three base exchanges code for leucine, arginine and serine. There are 17 serine residues conserved in both chains, 7 having identical codons, and 4 with just one base exchanged. Surprisingly, the remaining 6 serine codons have 2 or all bases converted to their complement. Such complementary base exchanges which also occur at other positions, have been designated with lowercase letters in fig.2.

4. DISCUSSION

Mouse α_1(IV) and α_2(IV) sequences are moderately conserved in the entire NC1 domain, the most highly conserved regions comprising the 12 cysteine residues. This homology is confined to the NC1 domain, since the main triple helices of both chains are not homologous except for the tripeptide repeats and the position of most of the triple-helical interruptions [9]. On the other hand, the sequences of α_1(IV) of mouse and human are highly conserved throughout the α_1-chain as far as both sequences are known. Only seven amino acid residues out of 229 differ between the α_1(IV) NC1 domain of mouse and human (cf. [3–5]). Of the 222 identical residues, 148 (67%) have also identical codons. Hence, the high degree of homology at the protein level is reflected at the DNA level as well. Many of the matching DNA strings have a length of 17 or 20 bp, two being even 28 and 32 bp long (cf. table 1). Thus, mouse cDNA clones have been used successfully to isolate corresponding human cDNA clones (cf. [4,5]). The DNA sequences of both mouse and human α_1(IV) NC1 domain are, to a similar extent, homologous to the corresponding mouse α_2(IV) sequence (cf. table 1). This may suggest that the α_1- and α_2-chains started to diverge before the separation of the genome of mouse and man.

The homology between the two NC1 domains of the different α(IV)-chains is nevertheless quite pronounced and similar to that found for the carboxypropeptides of collagen I [18]. We assume that there has been strong evolutionary pressure to conserve the structure of the NC1 domains as important aggregation and crosslinking sites for the type IV collagen network, while the flexibility of the network tolerated a much faster divergence of the sequences of both chains in the triple-helical domain, in contrast to the restrictions on divergence for the fibrillar collagen I [19]. The variations in structure of the NC1 domain of α_1(IV) and α_2(IV) may even be necessary for the correct selection of two α_1- and one α_2-chain to form a triple-helical molecule, as the NC1 domains are also believed to be involved in the correct alignment of the three chains before triple helix formation, in analogy to the C-terminal propeptides of collagen I [20].

While the carboxypropeptides of the interstitial collagens are cleaved off after triple helix forma-

tion, the NC1 domains are retained to serve as a crosslinking region for network formation. It has been a long-standing question whether at least some processing occurs at the C-terminal ends of the NC1 domains. Comparison of the C-terminal peptide sequences of both chains [21] with the cDNA sequences presented here and in [3–5] clearly shows that no processing takes place as the cDNA-derived protein sequences are identical to those of the native protein. The higher molecular mass for the two NC1 domains which has been observed intracellularly [22] may be due to incompletely folded and disulfide-linked NC1 domains.

Neither NC1 domain contains a carbohydrate attachment site Asn-X-Ser/Thr which is present in the carboxypeptides of the interstitial collagens. Therefore, they can contain only *O*-linked sugars. Thus, the glucosamine found earlier in preparations of murine NC1 (less than one residue per chain) [23] must result from a contamination.

ACKNOWLEDGEMENTS

This research was supported by the Deutsche Forschungsgemeinschaft (project Kü 70/21-1) and the Fritz Thyssen Stiftung.

REFERENCES

[1] Timpl, R., Wiedemann, H., Van Delden, V., Furthmayr, H. and Kühn, K. (1981) Eur. J. Biochem. 120, 203–211.
[2] Glanville, R.W., Quian, R.-Q., Siebold, B., Risteli, J. and Kühn, K. (1985) Eur. J. Biochem. 152, 213–219.
[3] Oberbäumer, I., Laurent, M., Schwarz, U., Sakurai, Y., Yamada, Y., Vogeli, G., Voss, T., Siebold, B., Glanville, R.G. and Kühn, K. (1985) Eur. J. Biochem. 147, 217–224.
[4] Pihlajaniemi, T., Tryggvason, K., Myers, J.C., Kurkinen, M., Lebo, R., Cheung, M.C., Prockop, D. and Boyd, C.D. (1985) J. Biol. Chem. 260, 7681–7687.
[5] Brinker, J.M., Gudas, L.J., Loidl, H.R., Wang, S.-Y., Rosenbloom, J., Kefalides, N.A. and Myers, J.C. (1985) Proc. Natl. Acad. Sci. USA 82, 3649–3653.
[6] Babel, W. and Glanville, R.W. (1984) Eur. J. Biochem. 143, 545–556.
[7] Schuppan, D., Glanville, R.W. and Timpl, R. (1982) Eur. J. Biochem. 123, 505–512.
[8] Schuppan, D., Glanville, R.W., Timpl, R., Dixit, S.N. and Kang, A.H. (1984) Biochem. J. 220, 227–233.
[9] Schwarz, U., Schuppan, D., Oberbäumer, I., Deutzmann, R., Timpl, R. and Kühn, K. (1986) Eur. J. Biochem. 157, 49–56.
[10] Lehman, J.M., Speers, W., Swartzendruber, D.E. and Pierce, G.B. (1974) J. Cell. Physiol. 84, 13–28.
[11] Okayama, H. and Berg, P. (1982) Mol. Cell. Biol. 2, 161–170.
[12] Vieira, J. and Messing, J. (1982) Gene 19, 259–268.
[13] Kieny, M.P., Lathe, R. and Lecocq, J.P. (1983) Gene 26, 91–99.
[14] Birnboim, H.C. and Doly, J. (1979) Nucleic Acids Res. 7, 1513–1522.
[15] Sanger, F., Coulson, A.R., Barrell, B.G., Smith, A.J.H. and Roe, B.A. (1980) J. Mol. Biol. 143, 161–178.
[16] Barker, W.C., Hunt, L.T., Orcutt, B.C., Gerorge, D.G., Veh, L.S., Chen, H.R., Blomquist, M.C., Johnson, G.C. and Dayhoff, M.O. (1983) in: Atlas of Protein Sequence and Structure, Protein Data Base, vol.7, National Biomedical Research Foundation, Washington, DC.
[17] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[18] Yamada, Y., Kühn, K. and Crombrugghe, B. (1983) Nucleic Acids Res. 11, 2733–2744.
[19] Hofmann, H., Fietzek, P. and Kühn, K. (1980) J. Mol. Biol. 141, 293–314.
[20] Doege, K.J. and Fessler, J.H. (1986) J. Biol. Chem. 261, 8924–8935.
[21] Siebold, B. (1986) Thesis, University of Munich.
[22] Timpl, R., Oberbäumer, I., Von der Mark, H., Bode, W., Wick, G., Weber, S. and Engel, J. (1986) Ann. NY Acad. Sci. 460, 58–72.
[23] Weber, S., Engel, J., Wiedemann, H., Glanville, R.W. and Timpl, R. (1984) Eur. J. Biochem. 139, 401–410.
[24] Oberbäumer, I. (1986) Gene, in press.